

Local Entropy Descriptor of Motion History Image for Human Action Recognition

De Zhang

Department of Automation, Beijing University of Civil Engineering and Architecture
ZhanLanGuan Road No. 1, Xicheng District, Beijing, China
zhangde@bucea.edu.cn

Abstract

Human action recognition is one of the most promising research areas at the moment. In this paper, we present an entropy based effective approach to recognize human activities from video sequences. Our approach provides a new description directly from Motion History Image (MHI) which is an efficient real-time representation for human action. Local entropy of pixels in MHI is computed to characterize the texture of motion filed. It helps to match the moment-based feature statistically. In an evaluation of this on well established benchmark data sets, we achieve high recognition rates. Using the popular Weizmann data set, we achieve the best accuracy of 98.9% in a leave-one-out cross validation procedure. Using the well-known KTH data set, we evaluate the robustness of the proposed approach over different scenarios. As a result, this new method is computationally efficient, robust with respect to appearance variation, and straight forward to be applied as it builds itself on well established and understandable concepts.

Keywords

Action Recognition; Local Entropy; Motion History Image

Introduction

In the computer vision communities, human action recognition in videos has been extensively studied and drew widespread attention in both academic and industrial research. It becomes a key component in many computer vision applications, such as video surveillance, human computer interaction, video games and multimedia retrieval, et al. Simply speaking, human action recognition attempts to automatically identify whether a person is walking, jumping, running or performing other types of action. It is a challenging problem because in practical scenarios obtaining a complete analytic description is extremely hard.

Most existing methods for human action recognition can be classified into two categories based on the frame-sequence descriptors used. One category of method is based on the sparse interest points [1-9]. The other is based on the dense holistic features [10-14, 28]. The information over space and time is combined to form a global representation, such as bag of words or a space-time volume. Then, a classifier is trained to label the resulting representation.

Both of these two kinds of descriptors have their limitations. The dense descriptor is not robust to variations in viewpoints and scales [15, 16]. So it doesn't model the actions well enough. However, the sparse descriptor doesn't emphasize on modeling the background but the background is also informative for the recognizing of human actions which contains contextual information for actions [3, 5]. Many interesting human activities are characterized by a complex temporal composition of simple actions. The Motion History Image (MHI) method recursively integrates the motion throughout a video action sequence into a single image to represent human activity. For recognition, higher-order moment features computed from the template were statistically matched to trained models [17, 18]. The approach was used within several interactive virtual environments requiring computer interpretation of the participant's motions. Although the MHI method has been proved to be effective in real-time constrained situations, it has yet a number of limitations related to the global image feature calculations and specific label-based recognition.

In this paper, we address these limitations by extending the approach with a new descriptor to computer a local

(dense) pixel entropy field directly from the MHI for capturing the texture information of the movement. Since MHI is not robust to appearance changes of human silhouette, a further description is usually required to not only capture mostly the motion information but also be insensitive to covariate condition changes that affect the appearance. Shannon Entropy [19] is presented to encode the randomness of pixel values in the silhouette images over a complete gait cycle. But it generally considers the values of probability between 0 and 1. To overcome this, local entropy is investigated which is one of the motivations of this work.

Based on computing local entropy, the new descriptor could render in a single image the randomness of pixel values in MHI which contains the raw motion information. Dynamic body areas which undergo consistent relative motion during one kind of action will lead to high entropy value. On the other hand, those areas that remain static would correspond to low values. A Local Entropy Descriptor (LEnD) thus captures mostly both global motion information and local texture appearance variation when performing an action. Compared with original MHI [17] and hierarchical MHI [20], LEnD keeps the strength of temporal history templates and the compensation for the varying speeds that are common to articulated human motion. In the meantime, it avoids the aforementioned weakness of existing representations.

We demonstrate the effectiveness of our LEnD representation through extensive experiments on the public benchmark data sets [21, 22, 27]. Our results suggest that the proposed LEnD of MHI outperforms other MHI related methods and also the baseline method. In the reminder of this paper, we first review the idea of temporal templates of MHI in next section. Then, we present the construction of Local Entropy Descriptor for MHI in the third section in detail. In the fourth section, we introduce the dataset used in the evaluation procedure and present the experimental results. Finally, we present our conclusions in the last section.

Motion History Image

The basic MHI framework is proposed for representing and recognizing human motions [17, 23]. This approach aims to extract recognizing information of holistic patterns of motion in an accumulating way. It thus gives a little care to structural features. The basis of this representation is a temporal template used for characterizing motion but originally is constrained to very particular domains, such as periodicity or facial motion. The general MHI template is targeted at representing arbitrary human movements. The advantage of the method is the use of a compact, yet descriptive, real-time representation capturing a sequence of motions in a single static image. The MHI is constructed by successively layering selected image regions over time using a simple update rule as in equation (1).

$$MHI_{\delta}(x, y) = \begin{cases} \tau & \text{if } \varphi(I(x, y)) \neq 0 \\ 0 & \text{else if } MHI_{\delta}(x, y) < \tau - \delta \end{cases} \quad (1)$$

where each pixel (x, y) in the MHI is marked with a current timestamp τ if the function φ signals object presence (or motion) in the current video image $I(x, y)$. The remaining timestamps in the MHI are removed if they are older than the decay value $\tau - \delta$. This update function is called for every new video frame analyzed in the sequence.

The function φ that selects a pixel location in the input image for inclusion into the MHI can be arbitrarily specified. Since the template representation captures both the position and temporal history of a moving object, many possibilities for selecting regions of interest are applicable. Detectors may include background subtraction, image differencing, optical flow, edges, stereo-depth silhouettes, flesh-colored regions, etc. With an object selection process for φ , the representation can accommodate slowly moving regions ($<1\text{pixel/frame}$) that would otherwise be missed by image differencing or standard optical flow. For the results presented here, we used a threshold image differencing method.

To illustrate the result of MHI construction, key frames from the sequences of different human actions performing in Weizmann data set and the corresponding (cumulative) MHIs are presented in Fig. 1. For the conveniences of displaying the timestamp pixel values in the templates are linearly mapped to graylevel values 0~255. The decay parameter δ in equation (1) is set to 10 (in frame numbers) for the illustration in Fig. 1. Here the brightness of a pixel corresponds to its recency in time. That is to say the brighter pixels are the most current timestamps.

Depending on the value chosen for the decay parameter δ , an MHI can represent a wide history of movement.

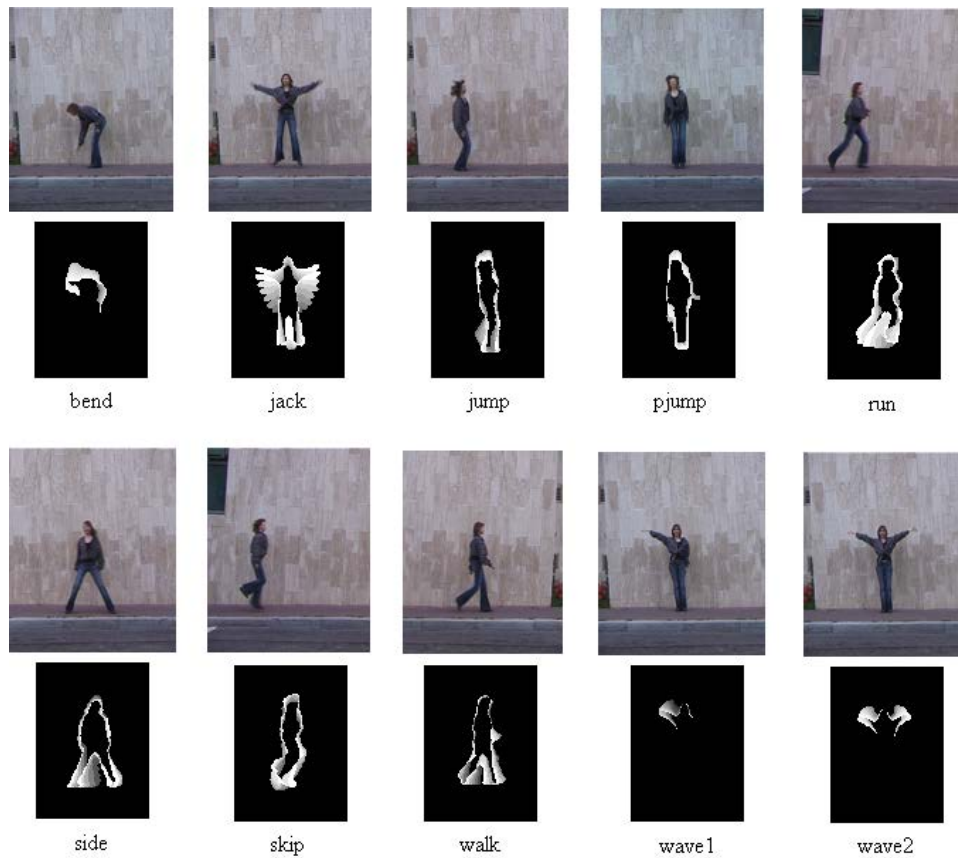


FIG. 1 EXAMPLES OF MHI (AT THE BOTTOM OF EACH PAIR) CORRESPONDING TO ORIGINAL FRAMES (AT THE TOP OF EACH PAIR) FROM VIDEOS OF DIFFERENT KINDS OF ACTIONS IN THE WEIZAMNN DATA SET

The initial application of MHIs to recognize human movements [23] was to extract several higher-order scale and translation invariant moment features and statistically match them to stored model examples using the Mahalanobis distance. The limitation with this method is the holistic appearance changes of the moment features calculated from the entire template, though it is successful in constrained situations with single and multiple cameras. Any occlusions of the body or errors from the implementation of φ would generate serious recognition failures. Also the recognition method was limited to only label-based recognition, where it could not give rise to any information other than specific identity matches. This motivated us to consider a more localized approach to motion analysis of the MHI.

Local Entropy Descriptor

According to equation (1), the MHI records the history impression of motion by layering the φ regions over time. It is quite apparent in such a temporal template, as shown in Fig. 1; and the progression of movements is captured in the dark-to-light intensity gradients. The unique motion patterns of each human action category thus result in its own texture style reflected through MHIs.

Shannon entropy is a statistical measure of randomness that can be used to characterize the texture of the input image. It measures the uncertainty associated with a random variable. Considering the intensity value of the silhouettes at a fixed pixel location as a discrete random variable, the entropy of this variable can be computed as in equation (2).

$$H(x, y) = -\sum_{k=1}^K p_k(x, y) \log_2 p_k(x, y) \quad (2)$$

where x, y are the pixel coordinates and $p_k(x, y)$ is the probability that the pixel takes on the k_{th} value. Pun [24] has

used Shannon's concept for defining the entropy of an image assuming that an image is entirely represented by its gray level histogram only. This approach using Shannon entropy function has led to successful segmentation algorithms. In this entropy, the measure of ignorance or the information gain is taken as $\log(1/p_i)$. But statistically ignorance can be better represented by $(1-p_i)$ when compared to $(1/p_i)$. Reyni entropy is an improvement over Shannon entropy. The exponential entropy function of Pal and Pal entropy is found to be more effective for the image specific applications as both the upper and the lower limits bounds of the exponential entropic measure are defined unlike the logarithmic measure which is undefined when either the occurrence probability of the event is zero or when the events are equally likely and the number of events is large. To overcome this problem, a new form of entropy calculation is presented in this work, which is termed as local entropy descriptor (LEnD).

The LEnD of an image outputs an entropy image of the same size as the input. Firstly, it computes the Shannon's entropy value of the neighborhood around each pixel in the input image. The entropy value is taken as the corresponding pixel's grayscale in the output image. The key point of this stage is to choose an appropriate size for the neighborhood, which is generally related with the size of input image. In this work, we set it as an 8-by-8 square area. Secondly, the LEnD should be normalized for the convenience of recognition step. The input MHIs usually doesn't share the same size of foreground region due to the multiformity of human activities. The uniform LEnD of MHIs is thus necessary for further handling with classification algorithms. Fig. 2 shows some examples of LEnD computed from MHIs of different actions in Weizmann data set.

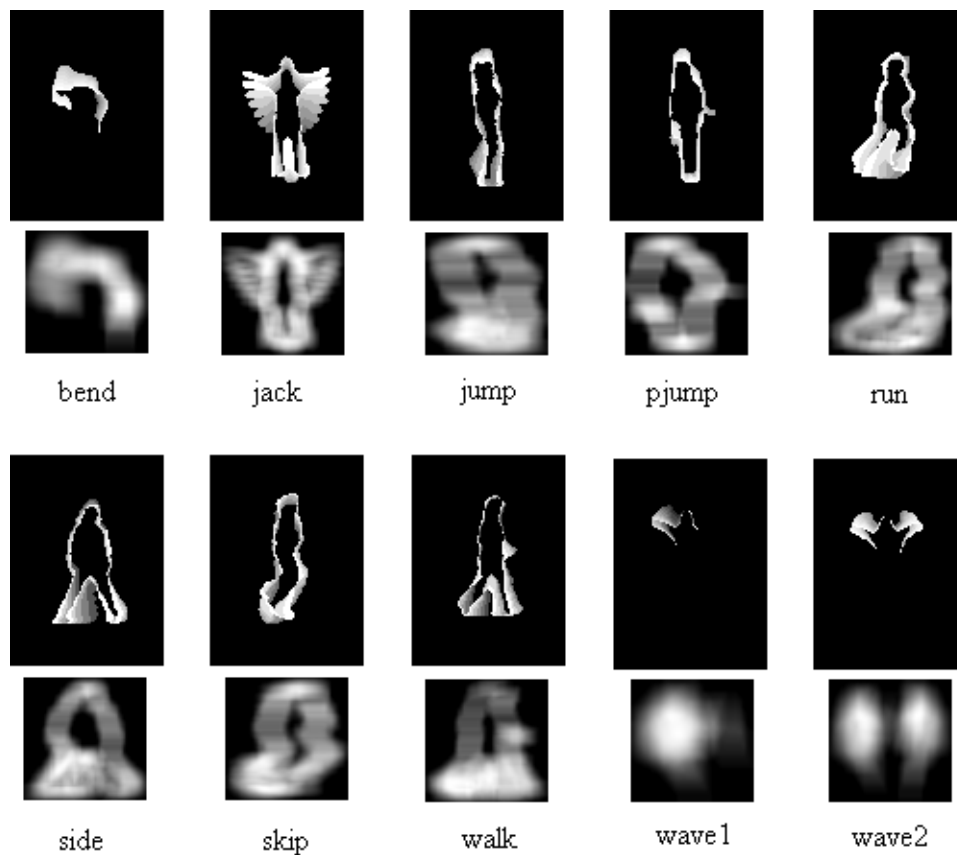


FIG. 2 EXAMPLES OF LEnD (AT THE BOTTOM OF EACH PAIR) CORRESPONDING TO MHIS (AT THE TOP OF EACH PAIR) CONSTRUCTED FROM DIFFERENT KINDS OF ACTIONS IN THE WEIZMANN DATA SET

It can be clearly seen that the texture pattern of MHIs is encoded with local entropy values in the LEnD images. These patterns could help a lot to decrease the intra-class difference, which is a challenging problem for human action recognition. Furthermore, most of temporal motion information remain in LEnD in order to provide significant inter-class variations. Note that in Fig. 2 only one frame extracted from an action video is illustrated. When applying in later experiments, for every action category there are several key frames combined to supply more discriminative power.

Experiments

Our evaluation experiments were performed on two most popular human action data sets, namely, the Weizmann data set [21, 22] (see Fig. 1), and the KTH data set [27] (see Fig. 5).

Weizmann Dataset

This dataset contains 90 videos with 10 different actions performed by 9 different actors. Fig. 1 shows some examples of these actions. Since the extracted silhouettes are also available with the original video sequences for the Weizmann dataset, we are free of background subtraction or other data pre-processing operations required. The silhouette masks provided consist of action samples including “run”, “walk”, “skip”, “bend”, “wave”, etc. All actions were recorded from the same view-point in a controlled environment.

According to the aforementioned technical approach, the temporal templates of MHI are constructed firstly for each action sequence. Then, the proposed LEnDs are calculated and normalized based on the extracted MHIs. During the calculation of LEnDs, we varied the number of key frames included in final feature extraction. In Fig. 3, we evaluate the sensitivity of LEnD to the number of key frames Num_{key} by plotting the classification error rate as a function of Num_{key} . The testing procedure was performed by using leave-one-out cross validation. That is, for every video sequence, we remove the entire of it from the database while other action sequences of the same person remain. When using the nearest neighbor classifier, the removed sequence is kept as a test sequence and all the remaining sequences were used as training samples.

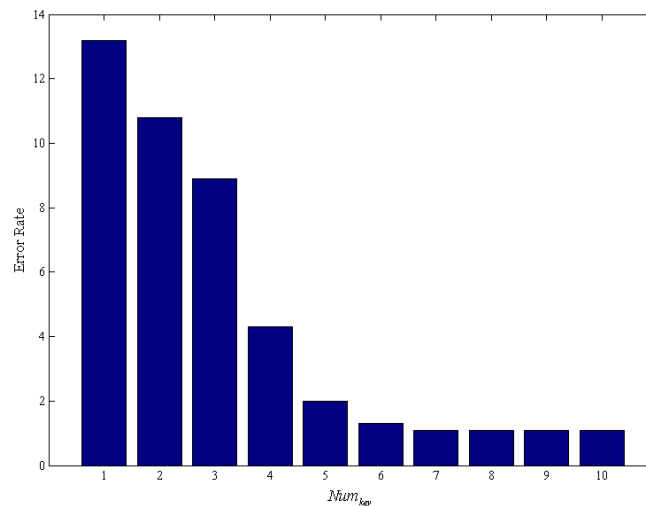


FIG. 3 EVALUATION OF LEND WITH VARYING KEY FRAME NUMBERS

As shown in Fig. 3, the LEnD method is related to the number of key frames chosen for feature extraction when Num_{key} is between 1 and 7. However, when selected frames are bigger than 7, the error rate tends to be stable. This result conveys a message similar to what is reported in by [12] which found that 1–7 frames are sufficient for basic action recognition.

According to the results shown in Fig. 3, we performed the classification experiments with the Num_{key} as 7. Under this situation, we obtained a mean classification accuracy of 98.9% for all nine actions. The confusion matrix for all actions is shown in Fig. 4. It can be observed that only 1 out of 90 videos was misclassified in these experiments. The only error was caused by the “run” action which was confused with the “skip”. This might be a reasonable confusion considering the small difference of texture pattern encoded in the LEnDs between MHIs of these actions.

In comparison with other feature descriptors, we performed another set of experiments with MHI [17], hierarchical MHI [20] and other silhouetted based representations [29–31]. Under the same leave-one-out validation and other parameter settings, the experimental results are obtained and listed in Table 1 for the comparison of our approach and established state-of-the-art approaches. As can be seen, the proposed LEnD method demonstrates the

superiority and could outperform state-of-the-art methods. This shows that constructing LEnD representations of MHI could provide reliable recognition rates in classifying human action.

	bend	jack	jump	pjump	run	side	skip	walk	wave1	wave2
bend	10									
jack		10								
jump			10							
pjump				10						
run					9		1			
side						10				
skip							10			
walk								10		
wave1									10	
wave2										10

FIG. 4 CONFUSION TABLE FOR WEIZMANN DATA SET (AVERAGE ACCURACY 98.9%)

TABLE 1 COMPARISON WITH OTHER METHODS ON WEIZMANN DATA SET

Method	Accuracy
MHI [17]	96.7%
hierachical MHI [20]	97.8%
method in [29]	97.8%
method in [30]	98.9%
method in [31]	96.8%
this paper	98.9%

KTH Dataset

This data set consists of 600 low-resolution (120×160 pixel) videos depicting 25 persons, performing six actions each [27]. These actions appearing in the data set are walking, jogging, running, boxing, hand waving, and hand clapping. Four different scenarios have been recorded: outdoors *s1*, outdoors with scale variation *s2*, outdoors with different clothes *s3*, and indoors *s4*, as illustrated in Fig. 5.



FIG. 5 VIDEO FRAMES OF KTH DATA SET FOR THE FOUR DIFFERENT SCENARIOS

In case of the KTH data set, we perform the suggested 2 tests: cross-scenario evaluation and leave-one-out evaluation on the whole data set.

For the first setup, we performed training on subset of $\{s1\}$ which contains video sequences from outdoors environment. Each of the remaining three subsets $\{s2\}$, $\{s3\}$, and $\{s4\}$ is taken as the test set respectively. Thus, we could analyze the influence of different scenarios. Since the proposed LEnD method is based on silhouette images, the scenario could impact the quality of silhouette. Due to shadows and imperfect subtraction with the background,

these silhouettes contain “leaks” and “intrusions”. For the training-test pairs of $\{s1, s2\}$, $\{s1, s3\}$, and $\{s1, s4\}$, we could evaluate the robustness of our approach to the varying of scale, clothes and environment. Also, we performed the experiment on subset $\{s1\}$ only by using leave-one-out cross validation to provide a baseline for the evaluation. As a result, we obtain an average classification accuracy of 95.3% for all six actions as the baseline. For the cross-scenario results, the average accuracies are 93.3% of $\{s1, s2\}$, 92.0% of $\{s1, s3\}$, and 94.7% of $\{s1, s4\}$. The corresponding confusion tables are shown in Fig. 6.

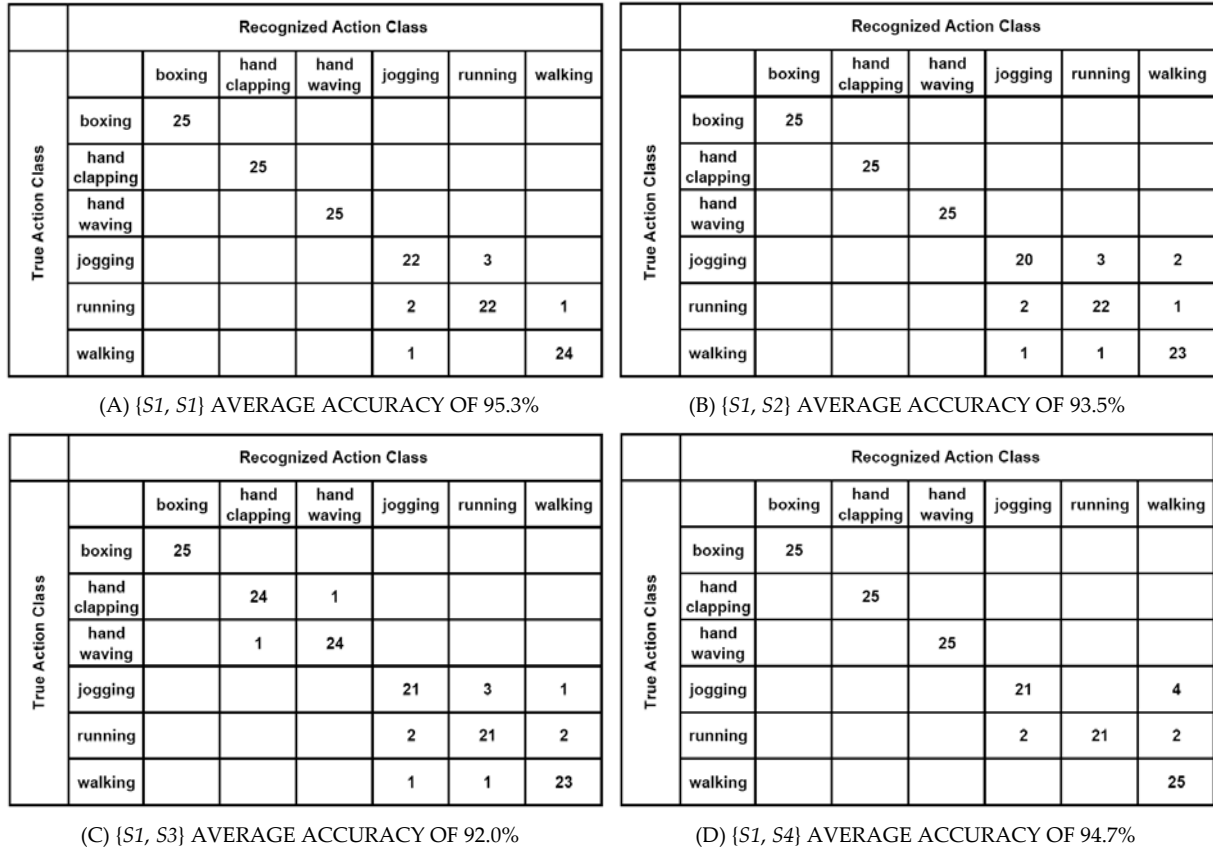


FIG. 6 CONFUSION TABLES FOR DIFFERENT TRAINING-TEST PAIRS

As can be seen in Fig. 6, the separation between leg actions and arm actions is clear and the two action classes that are the most difficult to be recognized are jogging and running. This result is similar to what is reported in [27]. Importantly, it can be observed that scenario differences would impact the classification rate, especially the clothes variation (Fig. 6(c)). Furthermore, Fig. 6(c) shows that a small confusion occurs between action classes waving and clapping, which doesn't happen in other cross-scenario situation. This is reasonable, since some types of clothes could occlude or follow person's hand motion or other kinds of actions. However, even in these cases, high classification rates are obtained by the proposed method. This shows that the proposed approach is not sensitive to most of scenario changes.

TABLE 2 RESULTS ON THE ENTIRE KTH DATA SET

Action Class	Accuracy
Boxing	96.2%
Hand clapping	93.1%
Hand waving	94.7%
Jogging	93.6%
Running	93.1%
Walking	93.7%

For the second setup, we use all scenarios to evaluate our approach. In the experiments, the classic leave-one-out training method, i.e. 24 persons for training and the other one for test, was adopted and we repeated the experiments 25 times to obtain the average results. Note that in one action class, each person has four video

sequences of different scenario. The results are shown in Table 2.

According to Table 2, the average accuracy is computed to be 94.1% which is better than the initial reported result on KTH data set in [27]. In addition, we performed experiments by using direct MHI and hierarchical MHI for the comparison. The corresponding average accuracies are 90.3% and 91.5%. Hence, all the experimental results from both data sets demonstrate the superiority and robustness of the proposed method.

Conclusions

In this paper we introduce a Local Entropy Descriptor (LEnD) which characterizes the local texture pattern of human actions by using the Motion History Image (MHI) as input. Using this descriptor for MHI, we extracted a set of dynamical and region metric invariants of the temporal templates of the dynamical system, and used it for action recognition. Experimental validation of the feasibility and potential merits of carrying out action recognition using this descriptor is demonstrated on real videos of human actions.

ACKNOWLEDGMENT

This work was supported by Beijing Natural Science Foundation (4144070 & 4144072).

REFERENCES

- [1] L. Cao, Z. Liu, and T. Huang, "Cross-dataset action detection," in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp: 1998-2005, 2010.
- [2] A. Kovashka and K. Grauman, "Learning a hierarchy of discriminative space-time neighborhood features for human action recognition," in Proc. of IEEE Conf. on Computer Vision and Pattern Recognition, pp: 2046-2053, 2010.
- [3] I. Laptev, "On space-time interest points," International Journal of Computer Vision, vol. 64, pp: 107-123, 2015.
- [4] I. Laptev, M. Marszalek, C. Schmid and B. Rozenfeld, "Learning realistic human actions from movies," in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp: 1-8, 2008.
- [5] R. Mattivi and L. Shao, "Human action recognition using LBP-TOP as sparse spatio-temporal feature descriptor," in Proceedings of International Conference on Computer Analysis of Images and Patterns, 2009.
- [6] J. C. Niebles, C. W. Chen, and L. Fei-Fei, "Modeling temporal structure of decomposable motion segments for activity classification," in European Conference on Computer Vision, vol. 6312, pp: 392-405, 2010.
- [7] K. Rapantzikos, Y. Avrithis, and S. Kollias, "Dense saliency based spatiotemporal feature points for action recognition," in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp: 1454-1461, 2009.
- [8] J. Sun, X. Wu, S. Yan, L. Cheong, T. Chua, and J. Li, "Hierarchical spatio-temporal context modeling for action recognition," in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp: 2004-2011, 2009.
- [9] A. Yao, J. Gall, and L. V. Gool, "A Hough transform-based voting framework for action recognition," in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp: 2061-2068, 2010.
- [10] N. Ikizler-Cinbis and S. Sclaroff, "Object, scene and actions: combining multiple features for human action recognition," in European Conference on Computer Vision, vol. 6311, pp: 494-507, 2010.
- [11] M. Lewandowski, D. Makris, and J. C. Nebel, "View and style-independent action manifolds for human activity recognition," in European Conference on Computer Vision, vol. 6316, pp: 492-505, 2010.
- [12] D. Weinland, M. Özuysal, and P. Fua, "Making action recognition robust to occlusions and viewpoint changes," in European Conference on Computer Vision, vol. 6313, pp: 635-648, 2010.
- [13] L. Yeffet and L. Wolf, "Local trinary patterns for human action recognition," in Proc. of IEEE International Conference on Computer Vision, pp: 492-497, 2009.
- [14] Z. Zeng and Q. Ji, "Knowledge based activity recognition with dynamic bayesian network," in European Conference on Computer Vision, vol. 6316, pp: 645-658, 2010.

- [15] A. Kläser, M. Marszalek, I. Laptev, and C. Schmid, "Will person detection help bag-of-features action recognition?" INRIA Technique report, 2010.
- [16] H. Wang, M. M. Ullah, A. Kläser, I. Laptev, and C. Schmid, "Evaluation of local spatio-temporal features for action recognition," in Proceedings of British Machine Vision Conference, pp: 32-26, 2009.
- [17] A. Bobick and J. Davis, "The recognition of human movement using temporal templates," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 23, issue. 3, pp: 257-267, 2001.
- [18] G. Bradski and J. Davis, "Motion segmentation and pose recognition with motion history gradients," in IEEE Workshop on Applications of Computer Vision, pp: 238-244, 2000.
- [19] K. Bashir, T. Xiang, and S. Gong, "Gait recognition using gait entropy image," in Proceedings of International Conference on Crime Detection and Prevention, pp: 1-6, 2009.
- [20] J. W. Davis, "Hierarchical motion history images for recognizing human motion," in IEEE Workshop on Detection and Recognition of Events in Video, pp: 39-46, 2001.
- [21] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri, "Action as space-time shapes," in Proceedings of International Conference on Computer Vision, vol. 2, pp: 1395-1402, 2005.
- [22] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri, "Action as space-time shapes," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 29, issue. 12, pp: 2247-2253, 2007.
- [23] J. Davis and A. Bobick, "The representation and recognition of action using temporal templates," in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp: 928-934, 1997.
- [24] T. Pun, "Entropic thresholding: a new approach," Computer Graphics, Vision and Image Processing, vol. 16, no. 3, 1981.
- [25] R. Alfred, "On measures of entropy and information," 4th Berkeley Symposium, vol. 1, pp: 547-561, 1961.
- [26] K. Schindler and L. V. Gool, "Action snippets: How many frames does human action recognition require?" in Proceedings of IEEE Confence on Computer Vision and Pattern Recognition, pp: 1-8, 2008.
- [27] C. Schuldt, I. Laptev, B. Caputo, "Recognizing human actions: a local SVM approach," in Proceedings of International Conference on Pattern Recognition, vol. 3, pp: 32-36, 2004.
- [28] N. Nikhil, L. Pedram, and S. S. Shankar, "Joint detection and recognition of human actions in wireless surveillance camera networks," in Proceedings of IEEE International Conference on Robotics and Automation, pp: 4747-4754, 2014.
- [29] L. Wang and D. Suter, "Recognizing human activities from silhouettes: Motion subspace and factorial discriminative graphical model," in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp: 1-8, 2007.
- [30] V. Kellokumpu, G. Zhao, and M. Pietikinen, "Human activity recognition using a dynamic texture based method," in Proceedings of British Machine Vision Conference, pp: 767-780, 2008.
- [31] L. Wang and C. Leckie, "Encoding actions via quantized vocabulary of averaged silhouettes," in Proceedings of International Conference on Pattern Recognition, pp: 3657-3660, 2010.



De Zhang was born in Hebei province, China on Nov. 2, 1979. He received the B. E. degree in automation from North China University of Technology, Beijing, China, in 2001, the second B. E. degree in computer science from Tsinghua University, Beijing, China, in 2003, and the Ph.D. degree in computer science from Beihang University, Beijing, China, in 2012.

He was with the Visualization and Intelligent Systems Laboratory, University of California, Riverside, U. S., from September 2009 to September 2010, as a Visiting Scholar. In July 2012, he joined the Department of Electrical and Information Engineering, Beijing University of Civil Engineering and Architecture, as a Faculty Member. His research interests include computer vision, pattern recognition, image processing, and machine learning.

Dr. Zhang is a member of China Computer Federation and Chinese Association of Automation. He is also a reviewer for several journals including Pattern Recognition, Neurocomputing, etc.